Determining the Optimal Process Technology for Performance-Constrained Circuits

Michael W. Boyer and Sudeep K. Ghosh
ECE 563 – Fall 2006
University of Virginia
<mwb7w,sudeep>@virginia.edu

ABSTRACT

The development of newer, smaller semiconductor processes enables significant gains in circuit performance over previous, larger processes. For applications which have no upper limit on their desired performance, the improved performance due to technology scaling is clearly beneficial. For applications which *do* have an upper limit on performance, however, it is not clear whether smaller technologies are actually desirable, since the improved performance generally comes at the cost of increases in leakage power, process variation, and power density.

We investigate this power tradeoff for a 16-bit carry bypass adder implemented in seven technologies: 1.6um, 0.6um, 130nm, 90nm, 65nm, 45nm, and 32nm. Our results show that the 0.6um technology dissipates less power than the smaller technologies for low duty cycle and low frequency operations, while the second most advanced 45nm technology dissipates the least power for high frequency, high duty cycle operations. Our analysis also raises a number of questions about the reliability of the transistor models for the advanced technologies. The results of this study indicate that designers of performance-constrained circuits should consider older process technologies when attempting to minimize power consumption.

1. INTRODUCTION

The semiconductor industry devotes enormous resources to continuing the reduction of transistor sizes through technology scaling. The most obvious goal of technology scaling, and the one made famous by Gordon Moore's eponymous law, is the doubling of transistor density. The other two main goals are the reduction of delay by 30%, which leads to an increase in clock speed of 43%, and the reduction of energy per transition by 65%, which leads to a power savings of 50% [1].

Technology scaling has been successfully meeting these goals for a number of years. Recently, however, factors that had previously been assumed insignificant have begun to have a large effect on the design and manufacturing of semiconductors. For example, the reduction in power due to scaling cited above ignores the impact of leakage currents. As the number of transistors on a die increases, the total amount of leakage current increases as well [1]. Additionally, the impact of process variation greatly increases as the size of transistors decreases. Another effect ignored by traditional scaling theory is increasing power density. As transistor sizes decrease, the number of transistors per unit area increases, leading to an increase in power density and possibly an increase in operating temperature. Since the magnitude of leakage currents increases with temperature, increasing power density can also lead to increased leakage power. Successfully scaling semiconductor technology is a much more complex process now than even a few years ago.

For many semiconductor devices, there is no upper limit on the desired rate of computation, so they are designed with the goal of maximizing performance while meeting certain area and power constraints. This is especially true of general-purpose microprocessors [2]. For products such as these, aggressive technology scaling certainly makes sense, since the previously stated goals of scaling lead directly to increased performance. Anecdotal confirmation of this can be seen in the aggressive pursuit of smaller and smaller technology nodes by both Intel and AMD, the two major general-purpose microprocessor manufacturers.

However, there also exists another class of devices, which we will call performance-constrained, for which there is an upper limit on the desired rate of computation and therefore no advantage in exceeding a certain baseline level of performance. A good example of this class would be a digital signal processor for which the maximum rate of incoming data is known a priori. Processing the data faster than it arrives will simply result in wasted processor cycles [2]. Devices such as these are designed with the goal of minimizing power and area while meeting a specific performance constraint. It is not clear whether technology scaling is beneficial in this case. The purpose of this study was to investigate this question further and determine if there are cases in which implementing a performance-constrained circuit in a larger, older technology would be more power-efficient than implementing the same circuit in a smaller, newer technology.

The rest of the paper is organized as follows: section two discusses related work, section three describes our experimental methodology, section four presents our results, and section five concludes.

2. RELATED WORK

There have been numerous articles published exploring the issue of low power circuit design, especially over the last decade. One of the first and most cited papers is [2]. Like most papers, [2] only attempts to minimize the power of a circuit implemented in a given technology, rather than considering the difference in power dissipation between technology nodes. Our assumption is that power-saving techniques like those discussed in [2] will have already been applied to a circuit before performing the type of analysis which we present.

[3] analyzes a sensor network processor implemented in three different technologies (1.5um, 250nm, and 130nm) to determine which technology dissipates the smallest amount of power across a wide range of activity factors and supply voltages. However, their study keeps the clock period fixed at 30 us, and thus their results are only relevant for circuits operating at this same clock rate.

3. METHODOLOGY

3.1 Circuit

For this study, we chose to analyze a 16-bit carry bypass adder with a block size of four. The four-bit adders were implemented using four one-bit mirror adders. The carry bypass adder was chosen because it is simple and easily understood but useful enough to be used in real applications. Using Cadence Virtuoso, we implemented the circuit in three different configurations, two using 1.6um and 0.6um technologies, both from AMI Semiconductor, and one using the Predictive Technology Models (PTM) developed at Arizona State University. We parameterized the lengths and widths of the transistors in the PTM circuit so that it could be easily used with all of the available PTM processes.

3.2 Predictive Technology Models

Predictive Technology Models (PTMs) are SPICE transistor models developed at Arizona State University and designed to model advanced process technologies [4]. The PTMs are based on ten primary process parameters that are considered most critical to technology scaling [4]. Unlike the earlier BPTMs developed at Berkeley, PTMs were explicitly designed to provide smooth scaling between processes, starting at 130nm and scaling all the way down to 32nm. Graduate students here at the University of Virginia have converted the SPICE PTMs into a format appropriate for use in the Cadence Spectre simulator.

The threshold voltages of the five PTMs are shown in Table 1 below. Note that as the technology gets smaller, the threshold voltage gets larger. This is opposite the trend that we would expect from traditional scaling theory [1]. And, oddly enough, these values are even quite different from those found in [4], the paper which asserts the validity of the PTMs. The PTM values found in [4] are all less than 0.3 V. Even more importantly, the PTM values below are significantly larger than the actual values reported by Fujitsu, Intel, IBM, TSMC, and TI for their own processes [4]. Unfortunately, these significant discrepancies in threshold voltages call into question the reliability of the PTMs, upon which much of our work is based.

Table 1: Threshold voltages for Predictive Technology Models

PTM	V_{th0n}	$\lvert V_{th0p} \rvert$
130nm	0.3782	0.321
90nm	0.397	0.339
65nm	0.423	0.365
45nm	0.466	0.4118
32nm	0.5088	0.450

3.3 Simulation

We simulated the adder circuit using Cadence Spectre for all seven technologies: the two AMI technologies and the five PTMs. For each technology, we varied the supply voltage and measured the resulting worst-case delay, the active power, and the leakage power. The delay was measured as the time from the inputs becoming valid to the last output (the 16th sum bit) becoming valid. The active power was measured as the average power dissipation during this same time interval. The leakage power was measured as the average of the power dissipation before and after this interval, in order to partially account for the dependence of leakage currents on input values. Note that active power as it is defined here is different than the standard definition of active power and actually includes some leakage power.

3.4 Total Power Calculation

We used the following equation from [3] to calculate the total power dissipation:

$$P_{total} = \alpha(T / T_{target})P_{active} + (1-\alpha(T / T_{target}))P_{leakage}$$

Here, α represents the duty cycle (the percent of clock cycles in which the adder is computing a result), T represents the delay through the adder, and T_{target} represents the maximum delay (the inverse of the desired frequency). For each combination of process technology, duty cycle, and frequency, we first computed the total power dissipation across all possible supply voltages. We then computed the minimum power among these supply voltages to determine the minimum power dissipation for that combination of technology, duty cycle, and frequency. We varied the frequency from one gigahertz down to ten kilohertz and duty cycle from 100% down to 0.00001%.

4. RESULTS

4.1 Delay

The effect of the supply voltage on the delay through the adder circuit for the PTMs is shown in Figure 1. As we would expect, the delay increases exponentially with decreasing supply voltage. Also, for most supply voltages, smaller technologies are faster. For a supply voltage of 0.4 V, however, the delay for the PTM nodes shows almost the opposite ordering, with the greatest delay through the 32nm circuit. Notice that a supply voltage of 0.4 V is actually less than the NMOS threshold voltage for the 32nm, 45nm, and 65nm nodes.

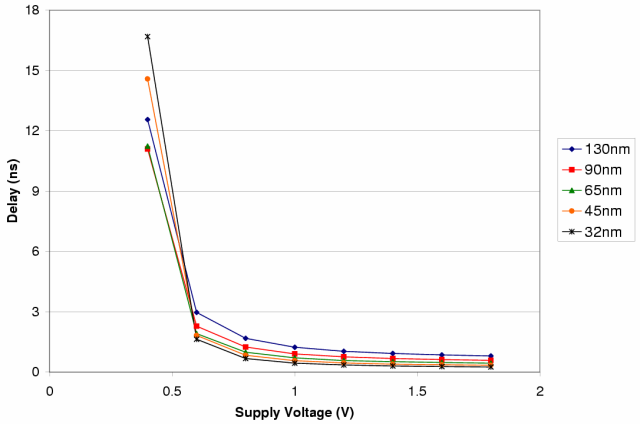

Figure 1: Adder delay for PTM nodes

The effect of the supply voltage on delay for the AMI nodes is shown in Figure 2. This graph is as we would expect, with delay increasing exponentially with decreasing supply voltage and the 1.6um circuit significantly slower than the 0.6um circuit for all supply voltages. For a supply voltage of 1.5 V, the 0.6um circuit is exceptionally slow but the 1.6um circuit does not even properly function (i.e., the delay is undefined).

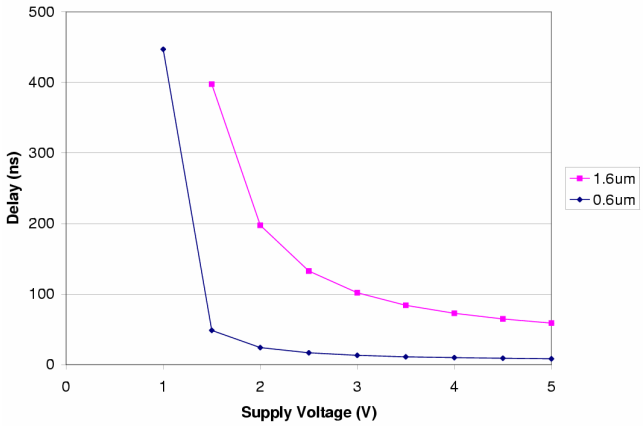

Figure 2: Adder delay for AMI nodes

4.2 Active Power

The impact of the supply voltage on the active power dissipation for the PTMs is shown in Figure 3. The overall trend of the graph is as expected, with power dissipation increasing as the supply voltage increases. At supply voltages below 1.4 V, the relationship among the curves is also exactly as we would expect, with power dissipation decreasing with reduced technology size. For supply voltages above 1.4 V, however, the 32nm circuit actually dissipates the most active power. Note that the power is shown in logarithmic units, so the increase in active power for 32nm is quite large.

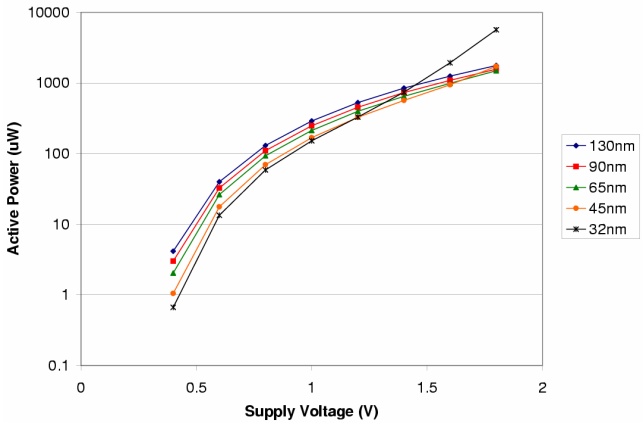

Figure 3: Active power dissipation for PTM nodes

The impact of supply voltage on the active power for the AMI nodes is shown in Figure 4. The overall trend of the two curves is as expected, with increasing supply voltage causing increased active power dissipation. The relationship between the two curves, however, is unexpected. Although we would expect a smaller technology to dissipate less active power, the 0.6um circuit actually dissipates more active power than the 1.6um circuit for all supply voltages.

4.3 Leakage Power

The effect of the supply voltage on the leakage power dissipation of the adder circuit for the PTMs is shown in Figure 5. Once again, the overall trend of the curves is as we would expect, with increasing supply voltage causing increasing leakage power. For supply voltages above 0.6 V, the relationship among the curves is also as we would expect, with smaller technologies dissipating more leakage power. At a supply voltage of 0.4 V, however, the relationship is essentially reversed, with the 45nm and 32nm circuits dissipating the least and second least leakage power, respectively, and the 130nm circuit actually dissipating the most leakage power.

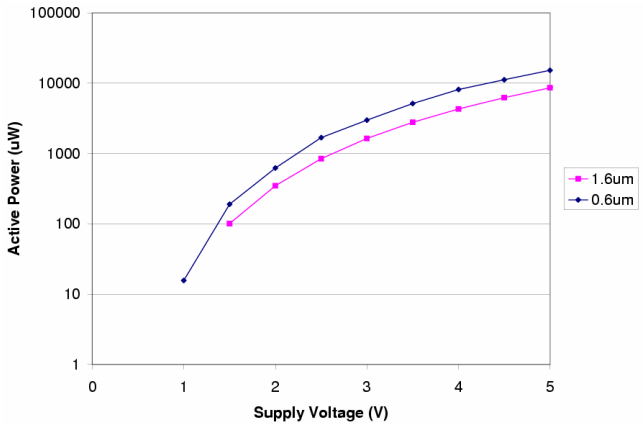

Figure 4: Active power dissipation for AMI nodes

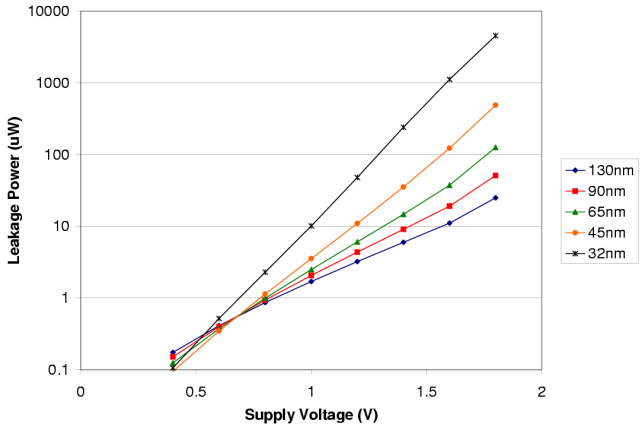

Figure 5: Leakage power dissipation for PTM nodes

The effect of supply voltage on the leakage power for the AMI nodes is shown in Figure 6. As with all of the previous graphs, the overall trend of the curves is as expected, with increasing supply voltages leading to increasing leakage power. The relationship between the curves, however, is extremely unexpected. Although we would anticipate a larger technology dissipating less leakage power, the 1.6um circuit actually dissipates approximately six orders of magnitude more leakage power than the 0.6um circuit. The 1.6um circuit even dissipates more than the PTM circuits for most supply voltages (although a direct comparison is not possible since the 1.6um circuit and the PTM circuits use much different ranges of supply voltages). This last discrepancy in particular leads to the conclusion that the power data derived from the 1.6um circuit is suspect, as will be shown even more clearly below.

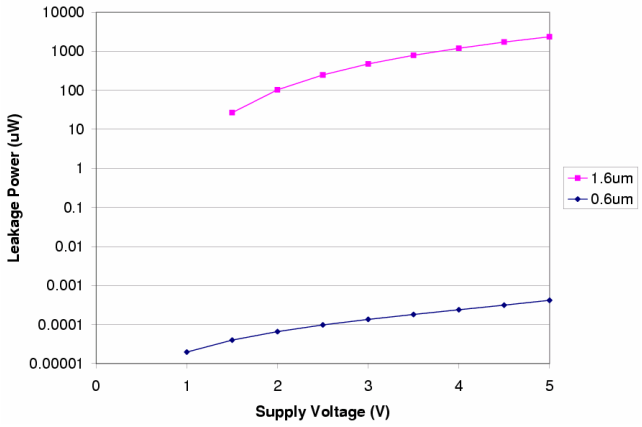

Figure 6: Leakage power dissipation for AMI nodes

4.4 Minimum Total Power

The minimum total power across all supply voltages for each combination of process technology, duty cycle, and desired frequency is shown in Figure 7. The circuits implemented in the two AMI technologies are too slow to properly function at extremely high frequencies. For all values of duty cycle and frequency for which the 1.6um technology can operate, it has the highest total power. This makes the 1.6um technology undesirable, from a power standpoint, for any potential circuit. We would have expected the 1.6um technology to have extremely low leakage power and therefore dissipate a small amount of power at low duty cycle and low frequency configurations; that this is not the case makes it likely that the 1.6um AMI model used in our simulations does not properly model power dissipation.

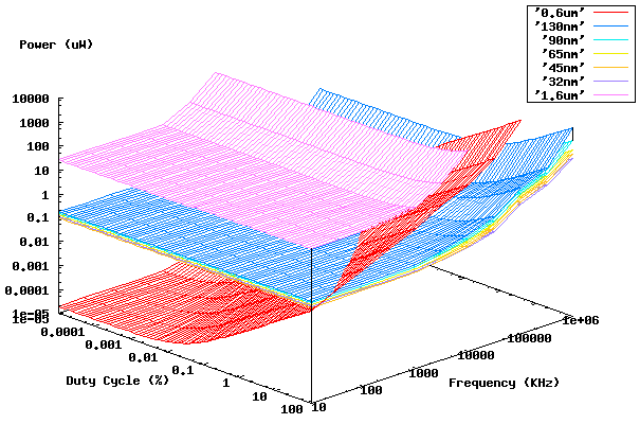

Figure 7: Minimum total power for all technologies

Since the 1.6um technology is uniformly the worst technology, Figure 8 presents the same minimum total power data as Figure 7 but with the 1.6um technology removed. At high frequency, high duty cycle operation, the PTMs dissipate less power than the 0.6um technology, due to their lower active power dissipation. At lower frequency and lower duty cycle operation, the 0.6um technology dissipates significantly less power than the PTMs, due to its lower leakage power dissipation.

There is very little differentiation between the total power dissipation of the different PTMs. From traditional scaling theory, we would expect the 32nm technology to be less power-efficient than all of the other PTMs in the leakage power dominant region, but it is actually the second most power-efficient of all of the PTMs in that region. The second smallest technology, 45nm, dissipates the least power of the PTMs in all regions except for the extremely low duty cycle and high frequency region.

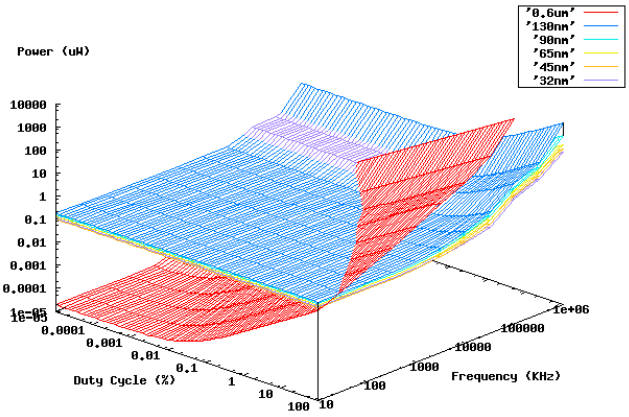

Figure 8: Minimum total power for PTMs and 0.6um

5. CONCLUSION

In this paper we have analyzed the power dissipation of seven different process technologies, from 1.6um to 32nm. Using this data, we computed the minimum total power dissipation for each technology across a wide range of duty cycles and frequencies. We found that the 0.6um technology dissipates the least amount of power for low duty cycle and low frequency operation, while smaller technologies dissipate less power at higher duty cycles and higher frequencies. Our simulations revealed a number of unexpected results for the PTMs as well as the 1.6um AMI technology model, which warrant further investigation.

It is worth noting that our analysis only considers one factor out of the many that would be involved in the choice of a process technology. In addition to power dissipation, a circuit designer would also want to consider: process variation and power density, which generally will favor larger technologies; area, which will favor smaller technologies; and fabrication cost. These factors tend to be straightforward to compare; power dissipation, as our results have shown, requires a more in-depth analysis to fully understand the tradeoffs between different process technologies.

6. REFERENCES

[1] Borkar, S. Design Challenges of Technology Scaling. *IEEE Micro, 19*, 4 (Jul-Aug 1999), 23-29.

[2] Chandrakasan, A. and Brodersen, R. Minimizing Power Consumption in Digital CMOS Circuits. *Proceedings of the IEEE, 83*, 4 (April 1995), 498-523.

[3] Hempstead, M., et al. An Ultra Low Power System Architecture for Sensor Network Applications. In *Proceedings of the 32nd International Symposium on Computer Architecture (ISCA '05)*, June 2005.

[4] Zhao, W. and Cao, Y. New generation of Predictive Technology Model for sub-45nm design exploration. In *Proceedings of the 2006 7th International Symposium on Quality Electronic Design (ISQED '06)*, 585-5